Topics in Applied Mathematics: Infinitely Large Neural Networks, 3341.751
E. Ryu
Spring 2022

Homework 6
Due 5pm, Friday, June 3, 2022

Problem 1: Let $\mathcal{X}$ be a nonempty set and let $\Theta \subseteq \mathbb{R}^P$. Let $f_\cdot(\cdot)\colon \Theta \times \mathcal{X} \to \mathbb{R}$ be a neural network and use the notation $f_\theta(x)$. Assume $\nabla_\theta f_\theta(x)$ is well defined for all θ and x and is continuous both in θ and x. Let $\theta_0 \in \Theta$ and define $h_\cdot(\cdot)\colon \Theta \times \mathcal{X} \to \mathbb{R}$ as

$$h_\theta(x) = f_{\theta_0}(x) + \langle \nabla_\theta f_{\theta_0}(x), \theta - \theta_0 \rangle_{\mathbb{R}^P}.$$

To clarify, $\nabla_\theta f_{\theta_0}(x) = (\nabla_\theta f_\theta(x))\,|_{\theta=\theta_0}$. So, $h_\theta(x)$ is the linearization of f_θ with respect to θ about θ_0. (Note, $h_\theta(x)$ is linear in θ, but nonlinear in x.) Define the PDK $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as

$$K(x, x') = \langle \nabla_\theta f_{\theta_0}(x), \nabla_\theta f_{\theta_0}(x') \rangle_{\mathbb{R}^P}, \qquad \forall\, x, x' \in \mathcal{X}.$$

Let $X_1, \dots, X_N \in \mathcal{X}$, and define $G \in \mathbb{R}^{N\times N}$ as

$$G_{ij} = K(X_i, X_j), \qquad \forall\, i, j \in \{1, \dots, N\}.$$

Assume G is strictly positive definite. Let

$$\Phi = \begin{bmatrix} (\nabla_\theta f_{\theta_0}(X_1))^\intercal \\ (\nabla_\theta f_{\theta_0}(X_2))^\intercal \\ \vdots \\ (\nabla_\theta f_{\theta_0}(X_N))^\intercal \end{bmatrix} \in \mathbb{R}^{N\times P}, \qquad \Delta = \begin{bmatrix} f_\star(X_1) - f_{\theta_0}(X_1) \\ f_\star(X_2) - f_{\theta_0}(X_2) \\ \vdots \\ f_\star(X_N) - f_{\theta_0}(X_N) \end{bmatrix} \in \mathbb{R}^N.$$

Consider the regression problem

$$\underset{\theta\in\mathbb{R}^P}{\text{minimize}} \quad \sum_{i=1}^N (h_\theta(X_i) - f_\star(X_i))^2.$$

Show that

$$\theta_\star = \theta_0 + \Phi^\intercal G^{-1}\Delta$$

is an optimal solution and that

$$h_{\theta_\star}(x) = f_{\theta_0}(x) + \sum_{j=1}^N K(x, X_j)(G^{-1}\Delta)_j, \qquad \forall\, x \in \mathcal{X}.$$

Remark. $\theta_\star$ is not the unique solution, but it is the so-called "minimum-norm" solution.

Remark. This problem considers learning with h_θ, the linearization of f_θ, rather than the actual neural network f_θ. Interestingly, the learned $h_{\theta_\star}$ is identical to the prediction function obtained via the NTK theory, which characterizes the training f_θ in the infinite-width limit. In fact, K is the neural tangent kernel of f_θ at $\theta = \theta_0$.

Problem 2: *NTK of random feature learning.* Consider the 2-layer MLP

$$f_\theta(x) = \sum_{i=1}^{M} \frac{1}{\sqrt{M}} \theta_i \sigma(a_i^\intercal x + b_i),$$

where $\sigma\colon \mathbb{R} \to \mathbb{R}$ is a continuous activation function, $a_1, \dots, a_N \in \mathbb{R}^d$ and $b_1, \dots, b_N \in \mathbb{R}$ are initialized as

$$(a_i)_j \sim \mathcal{N}(0, 1/d), \qquad b_i \sim \mathcal{N}(0, 1)$$

and not trained, and $\theta_1, \dots, \theta_M \in \mathbb{R}$ are trainable parameters. (So we assume f_θ outputs a scalar.) Let P be a probability measure with finite support. Consider training through

$$\underset{\theta \in \mathbb{R}^M}{\text{minimize}} \quad R[f_\theta],$$

and assume the risk $R\colon L^2(P) \to \mathbb{R}$ is Fréchet differentiable. Show that the gradient flow dynamics on the parameters

$$\frac{d\theta}{dt} = -\nabla_\theta R[f_\theta]$$

induces the dynamics

$$\frac{d}{dt} f_\theta = -L_\Theta[\partial_f R],$$

with

$$\Theta(x, x') = \frac{1}{M} \sum_{i=1}^{M} \sigma(a_i^\intercal x + b_i) \sigma(a_i^\intercal x' + b_i).$$

(Note, Θ is time-independent.) Also show that

$$\Theta \to \tilde{\Sigma}^{(2)}$$

in probability as $M \to \infty$ pointwise for inputs (x, x'), where

$$\Sigma^{(1)}(x, x') = \frac{1}{d} x^\intercal x' + 1.$$

and

$$\tilde{\Sigma}^{(2)}(x, x') = \mathbb{E}_{f \sim \mathcal{GP}(0, \Sigma^{(1)})}[\sigma(f(x)) \sigma(f(x'))]$$

Clarification. In the NNGP and NTK lectures, we used the variance parameters σ_A and σ_b. Here, we set $\sigma_A = \sigma_b = 1$ for the sake of simplicity.

Problem 3: *NTK with standard parameterization.* Consider the depth-2 MLP

$$\begin{aligned} f_\theta(x) &= y_2 \\ y_2 &= z_2, \qquad & z_2 &= A_2 y_1 + b_2 \in \mathbb{R}^{n_2}, \\ y_1 &= \sigma(z_1), \qquad & z_1 &= A_1 x + b_1 \in \mathbb{R}^{n_1}, \end{aligned}$$

where $x \in \mathbb{R}^{n_0}$, $A_\ell \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$, and $b_\ell \in \mathbb{R}^{n_\ell}$. Initialize the weights with

$$(A_1)_{ij} \sim \mathcal{N}(0, 1/n_0), \qquad (b_1)_i \sim \mathcal{N}(0,1)$$

and

$$(A_2)_{ij} \sim \mathcal{N}(0, 1/n_1), \qquad (b_2)_i \sim \mathcal{N}(0,1).$$

Consider training through

$$\underset{\theta}{\text{minimize}} \quad R[f_\theta],$$

and assume the risk $R\colon L^2(P) \to \mathbb{R}$ is Fréchet differentiable. For $n_1 < \infty$, the gradient flow dynamics

$$\frac{d\theta}{dt} = -\frac{1}{n_1} \nabla_\theta R[f_\theta]$$

induces the dynamics

$$\frac{d}{dt} f_\theta = -L_{\frac{1}{n_1}\Theta_t}[\partial_f R].$$

Find a formula for the NTK Θ_t and show that

$$\frac{1}{n_1}\Theta_0 \to \tilde{\Sigma}^{(2)} \otimes I_{n_2}$$

in probability as $n_1 \to \infty$ pointwise for inputs (x, x') at time $t = 0$, where $\tilde{\Sigma}^{(2)}$ is as defined in Problem 2.

Problem 4: *Gluing Lemma.* Let $\Theta \subseteq \mathbb{R}^d$ be nonempty. For any $\rho_1, \rho_2 \in \mathcal{P}(\Theta)$, define

$$\Pi(\rho_1, \rho_2) = \{\pi \in \mathcal{P}(\Theta \times \Theta) \,|\, \text{probability measures on } \Theta \times \Theta \text{ with marginals } \rho_1 \text{ and } \rho_2\}.$$

Let $\lambda, \mu, \nu \in \mathcal{P}(\Theta)$ and $\pi_{1,2} \in \Pi(\lambda, \mu)$ and $\pi_{2,3} \in \Pi(\mu, \nu)$. Define $P_i\colon \Theta \times \Theta \times \Theta \to \Theta$ for $i = 1, 2, 3$ as

$$P_1(\theta_1, \theta_2, \theta_3) = \theta_1, \qquad P_2(\theta_1, \theta_2, \theta_3) = \theta_2, \qquad P_3(\theta_1, \theta_2, \theta_3) = \theta_3.$$

Define $P_{i,j}\colon \Theta \times \Theta \times \Theta \to \Theta \times \Theta$ with $1 \leq i < j \leq 3$ as

$$P_{i,j}(\theta_1, \theta_2, \theta_3) = (\theta_i, \theta_j).$$

Show that there is a $\pi_{1,2,3} \in \mathcal{P}(\Theta \times \Theta \times \Theta)$ such that

$$P_{1\#}\pi_{1,2,3} = \lambda, \qquad P_{2\#}\pi_{1,2,3} = \mu, \qquad P_{3\#}\pi_{1,2,3} = \nu$$

and

$$\pi_{1,2} = P_{1,2\#}\pi_{1,2,3}, \qquad \pi_{2,3} = P_{2,3\#}\pi_{1,2,3}, \qquad \pi_{1,3} := P_{1,3\#}\pi_{1,2,3} \in \Pi(\lambda, \nu).$$

Hint. Disintegrate $\pi_{1,2}$ as

$$d\pi_{1,2}(\theta_1, \theta_2) = d\tilde{\mu}_{\theta_1}(\theta_2) d\lambda(\theta_1)$$

and $\pi_{2,3}$ as

$$d\pi_{2,3}(\theta_2, \theta_3) = d\tilde{\nu}_{\theta_2}(\theta_3) d\mu(\theta_2).$$

Define $\pi_{1,2,3}$ as

$$d\pi_{1,2,3} = d\tilde{\nu}_{\theta_2}(\theta_3) d\tilde{\mu}_{\theta_1}(\theta_2) d\lambda(\theta_1).$$

Problem 5: *Triangle inequality of the Wasserstein distance.* Let $\Theta = \Phi \subseteq \mathbb{R}^d$ and $p \in [1, \infty)$. Show that

$$W_p(\lambda, \nu) \leq W_p(\lambda, \mu) + W_p(\mu, \nu), \qquad \forall\, \lambda, \mu, \nu \in \mathcal{P}^p(\Theta).$$

Hint. Let $\pi_{1,2}$ and $\pi_{2,3}$ be feasible joint probability measures for the optimization problems defining $W_p(\lambda, \mu)$ and $W_p(\mu, \nu)$. (Do not assume $\pi_{1,2}$ and $\pi_{2,3}$ are optimal, since we do not know whether the minimuma are attained.) Using Problem 4, glue $\pi_{1,2}$ and $\pi_{2,3}$ to get $\pi_{1,2,3}$ and $\pi_{1,3}$. Finally, use the Minkowski inequality in $L^p(\pi_{1,2,3})$.

Problem 6: *Optimum of book shifting via duality.* Let $\Theta = \Phi = \mathbb{R}$, $c(\theta, \phi) = \|\theta - \phi\|$, and

$$\mu = \frac{1}{N} \sum_{i=1}^{N} \delta_i, \qquad \nu = \frac{1}{N} \sum_{i=1}^{N} \delta_{i+1}.$$

Show that $W_1(\mu, \nu) \geq 1$ by finding a suitable feasible $\varphi \in \mathcal{L}_1$ for the Kantorovich–Rubinstein dual

$$W_1(\mu, \nu) = \left(\begin{array}{ll} \underset{\varphi \in \mathcal{C}_0(\Theta)}{\text{maximize}} & \int_{\mathbb{R}} \varphi(\theta)\, d\mu(\theta) - \int_{\mathbb{R}} \varphi(\phi)\, d\nu(\phi) \\ \text{subject to} & \varphi \in \mathcal{L}_1 \end{array} \right).$$

Problem 7: Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. Let

$$\mathbb{R}^n_+ = \{(x_1, \dots, x_n) \in \mathbb{R}^n \mid x_1 \geq 0, \dots, x_n \geq 0\}$$

be the nonnegative orthant in $\mathbb{R}^n$. Consider the optimization problem

$$\begin{array}{ll} \underset{\theta \in \mathbb{R}^n}{\text{minimize}} & f(x) \\ \text{subject to} & x \in \mathbb{R}^n_+ \end{array}$$

and let $x^\star \in \mathbb{R}^n_+$ be an optimal solution. Show that

$$\frac{\partial f}{\partial x_i}(x^\star) \geq 0, \qquad \forall\, i = 1, \dots, n$$

and

$$\frac{\partial f}{\partial x_i}(x^\star) = 0, \qquad \forall\, i \text{ such that } x_i^\star > 0.$$

Problem 8: Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. Let

$$\Delta^n = \{(x_1, \dots, x_n) \in \mathbb{R}^n \mid x_1 + \cdots + x_n = 1,\, x_1 \geq 0, \dots, x_n \geq 0\}$$

be the probability simplex in $\mathbb{R}^n$. Consider the optimization problem

$$\begin{array}{ll} \underset{\theta \in \mathbb{R}^n}{\text{minimize}} & f(x) \\ \text{subject to} & x \in \Delta^n \end{array}$$

and let $x^\star \in \Delta^n$ be an optimal solution. Let

$$c = \min_{i=1,\dots,n} \frac{\partial f}{\partial x_i}(x^\star).$$

Show that

$$\frac{\partial f}{\partial x_i}(x^\star) = c, \qquad \forall\, i \text{ such that } x_i^\star > 0.$$

Problem 9: Let $f\colon \mathbb{R}^n \to \mathbb{R}$ and $k > 0$. Assume f is nonnegative homogeneous of degree k, i.e.,

$$f(\alpha x) = \alpha^k f(x), \qquad \forall\, \alpha \geq 0,\, x \in \mathbb{R}^n.$$

Assume f is differentiable at x_0. Show that (i)

$$\langle x_0, \nabla f(x_0) \rangle = k f(x_0)$$

(ii) and

$$\nabla f(\alpha x_0) = \alpha^{k-1} \nabla f(x_0), \qquad \forall\, \alpha > 0.$$

Hint. For (i), differentiate both sides of $f(\alpha x_0) = \alpha^k f(x_0)$ with respect to α and plug in $\alpha = 1$. For (ii), differentiate both sides of $f(\alpha(x_0 + te_i)) = \alpha^k f(x_0 + te_i)$ with respect to t and plug in $t = 0$.

Problem 10: Let $\sigma\colon \mathbb{R} \to \mathbb{R}$ defined as

$$\sigma(r) = \max\{r, 0\}$$

be the ReLU activation function. Of course, σ is nonnegative homogeneous of degree 1. Let $x \in \mathbb{R}^d$ and $\theta = (u, a, b) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}$. Define

$$f(\theta) = u\sigma(a^\intercal x + b).$$

Show that (i) $f(\theta)$ is nonnegative homogeneous of degree 2 and (ii) $f(\theta)$, is differentiable for (Lesbesgue) almost all $\theta \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}$.

Clarification. We view x as a fixed input.